The self-organization of genomes

Ramon Ferrer-i-Cancho [a,*]

[a] *Departament de Llenguatges i Sistemes Informàtics, TALP Research Center,*
Universitat Politècnica de Catalunya,
Campus Nord, Edifici Ω, Jordi Girona Salgado 1-3. 08034 Barcelona, Spain.

Núria Forns [b,c]

[b] *Departament de Física Fonamental, Universitat de Barcelona.*
Martí i Franquès 1, 08028 Barcelona, Spain.
Current address:

[c] *Networking Research Center on Bioengineering, Biomaterials and Nanomedicine*
(CIBER-BBN), Barcelona, Spain.

Abstract

Menzerath-Altmann law is a general law of human language stating, for instance, that the longer a word, the shorter its syllables. With the metaphor that genomes are words and chromosomes are syllables, we examine if genomes also obey the law. We find that longer genomes tend to be made of smaller chromosomes in organisms from three different kingdoms: fungi, plants and animals. Our findings suggest that genomes self-organize under principles similar to those of human language.

* Corresponding author. Phone: +34 93 4137870. Fax: +34 93 4137787.
Email address: ramon.ferrericancho@gmail.com (Ramon Ferrer-i-Cancho).

Kewords: Menzerath-Altmann law; genomes; chromosomes; self-organization; quantitative linguistics

Short communication

Human language and genomes are apparently very different information coding systems. However, various linguistic traditions are providing deeper insights into the complexity of genetic sequences [1,2] and it has been argued that both language and genomes have followed similar evolutionary strategies to attain higher complexity [3]. Here we will explore the similarities between language and genomes from a quantitative linguistics perspective. In particular, we will investigate if Menzerath-Altmann law, a well known law in quantitative linguistics [4,5], holds not only in human language but also qualitatively in genomes.

Menzerath-Altmann law is a universal law of languages [6,4,5] that states that "the longer a language construct the shorter its components (constituents)" [4]. If we take a word as construct and its syllables as components, the instantiation of the law gives, for instance, "the longer a word, the shorter its syllables" [4]. The length of a word can be measured in syllables and the length of a syllable can be measured in phonemes [4]. The law bears the name of the person who proposed the law qualitatively, i.e. Paul Menzerath [6], and that of the person who put it in mathematical form, i.e. Gabriel Altmann [4]. The law is regarded as evidence of self-organization in languages: e.g., the lengthening of a word must be compensated by the shortening of its syllables [7,5].

Here we aim to investigate if the law holds qualitatively in other information coding systems such as the genomes. Imagine that a genome is a word and its

chromosomes are syllables. Taking the genomes of many organisms, we will test if the longer the genome, the shorter its chromosomes. More precisely, we will study the relationship between the size of a genome in chromosomes (L_g) and the mean length of its chromosomes in million base pairs (L_c) in three kingdoms: animals, plants and fungi (see the Appendix for details about the data). We study this relationship qualitatively in the sense that we are only concerned about finding statistically significant negative correlations between L_g and L_c, although exact functions for fitting the dependency between the size of a construct and the size of its components have been proposed [4,5]. The rationale of this simple approach is the quest for the largest set of groups of organisms agreeing qualitatively with the law, neglecting, for the present study (a) the particular differences on the shape of this trend for different groups or organisms and (b) the possibility that this correlation might be significant but no candidate function yields a satisfactory fit.

Fig. 1 shows L_c versus L_g for four different groups of organisms. In order to assess whether there is a negative correlation between L_g and L_c, we employed a one sided Spearman rank correlation test [8]. A statistically significant negative correlation is found for fungi, angiosperm plants, cartilaginous fishes, jawless fishes, reptiles, birds and mammals, which is consistent with Menzerath-Altmann law qualitatively, but not in gymnosperm plants and ray-finned fishes (Table 1). Exceptionally, amphibians show a weak positive statistically significant correlation.

The fact that statistically correlations are not found for gymnosperm plants and ray-finned fishes does not imply that the Menzerath-Altmann law does not hold qualitatively within these groups. Indeed, statistically significant correlations are found for subgroups, e.g., ray-finned fishes from the genus

Oncorhynchus and from the genus *Corydoras*, as well as gymnosperm plants excluding the genus *Pinus* (Table 1).

In sum, the genomes with more chromosomes tend to be made of shorter chromosomes in various groups of organisms as words with more syllables tend to be made of shorter syllables in languages. As far as we know, the only previous evidence of Menzerath-Altmann law is only in the genomes of ants [2]. Our findings suggest that self-organization in the sense of organization without global or external control [9] is a general principle of genomes. Organisms within the same group obey the law autonomously, without any group director. Finally, our results suggest that human language and genomes share similar principles of self-organization.

References

[1] D. B. Searls, *Nature* **420**, 211 (2002).

[2] J. Wilde, H. Schwibbe, *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, G. Altmann, M. H. Schwibbe, eds. (Olms, Hildesheim, 1989), pp. 92–107.

[3] H.-Y. Zhang, *EMBO reports* **7**, 748 (2006).

[4] G. Altmann, *Glottometrika 2* **2**, 1 (1980).

[5] L. Hřebíček, *Quantitative Linguistics* **56** (1995).

[6] P. Menzerath, *Die Architektonik des deutschen Wortschatzes* (Dümmler, Bonn, 1954).

[7] R. Köhler, *Zur Linguistischen Synergetik: Struktur und Dynamik der Lexik* (Brockmeyer, Bochum, 1986).

[8] W. J. Conover, *Practical nonparemetric statistics* (Wiley, New York, 1999). 3rd edition.

[9] L. Steels, *Machine Intelligence* **15**, 205 (1996).

[10] J. Dolezel, J. Bartos, H. Voglmayr, J. Greilhuber, *Cytometry* **51**, 127 (2003).

Aknowledgements

We thank D. Lusseau for advice on statistical analysis R. Köhler for making us aware of previous work and C. Heidipriem for translation from German. This work was funded by a Juan de la Cierva contract from the Spanish Ministry of Education and Science under the projects BFM2003-08258-C02-02 and FIS2006-13321-C02-01 (RFC). It is strictly prohibited to use, investigate or develop, in a direct or indirect way, any of the scientific contributions of the author(s) contained in this work by any army or armed group in the world, for military purposes and for any other use which is against human rights or the environment, unless a written consent of all the persons in the world is obtained.

Appendix

The data were obtained from three public databases:

- Animals: Gregory, T.R. (2007). Animal Genome Size Database. http://www.genomesize.com.
- Plants: Bennett M.D. and Leitch I.J. 2005. Plant DNA C-values. Database (release 4.0, Oct. 2005). http://www.kew.org/cval/homepage.html. With the permission of the Trustees of the Royal Botanic Gardens, Kew.

- Fungi: Kullman, B., Tamm, H. and Kullman, K. 2005. Fungal Genome Size Database. http://www.zbi.ee/fungal-genomesize.

If there was more than one entry for a given species in a database, we applied the following criteria:

- If the number of chromosomes was the same in all entries, then the genome size was averaged.
- If the number of chromosomes was different, then the species was removed.

In many species, the number of chromosomes or the genome length in Million base pairs (Mb) was missing. For our analysis, the genome length in Mb is measured using the DNA C-value ($1C$). Accordingly, the number of chromosomes is measured using $1n$. Concerning the animal database, it was necessary to convert picograms (pg) of DNA to Mb of DNA through the formula [10]

$$\text{number of base pairs (in } Mb) = \text{mass (in } pg) \times 978. \tag{1}$$

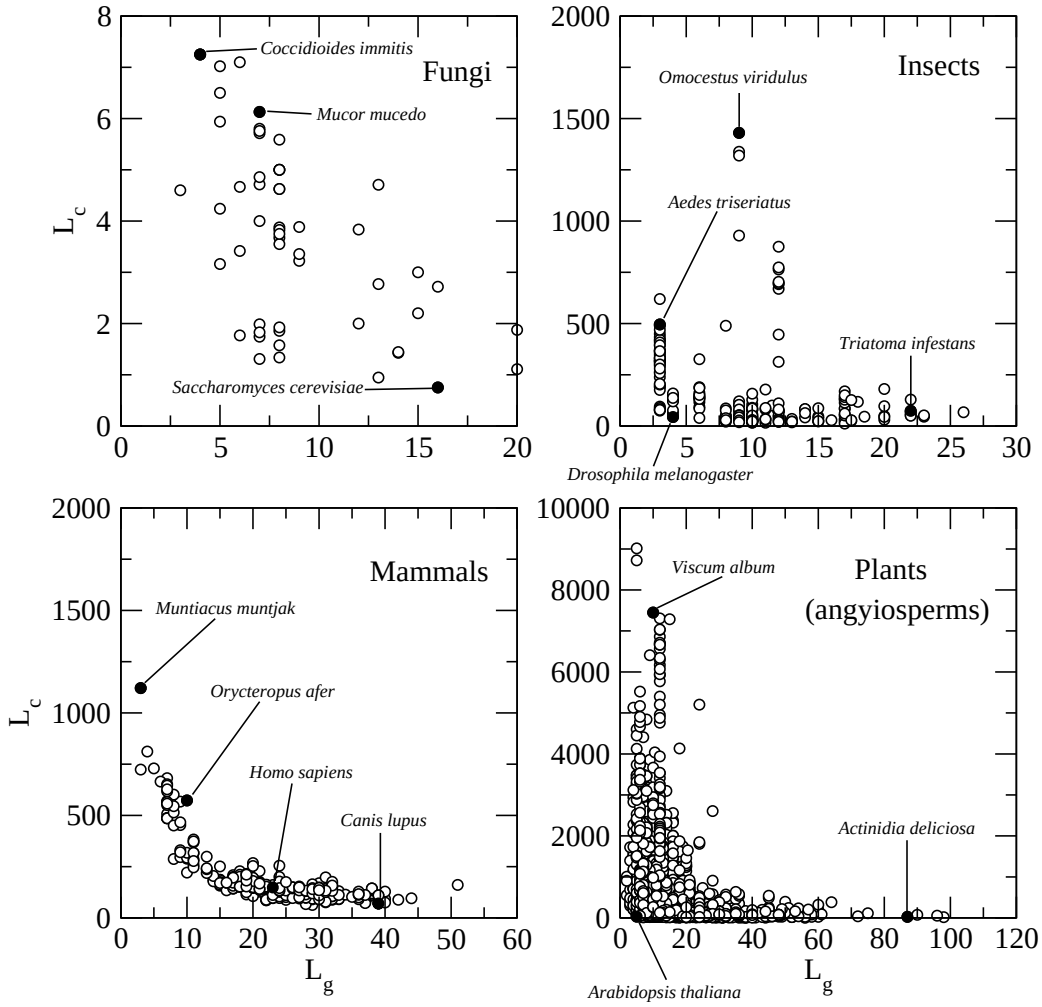

Fig. 1. The relationship between the size of a genome in chromosomes (L_g) and the mean length of its chromosomes in million base pairs (L_c) for four different groups of organisms. The points of some selected species are indicated. Some species have a common name. Regarding insects: *Drosophila melanogaster* (fruit fly), *Omocestus viridulus* (common green grasshopper) and *Aedes triseriatus* (eastern treehole mosquito). Regarding mammals: *Canis lupus* (wolf), *Homo sapiens* (modern human), *Orycteropus afer* (aardvark) and *Muntiacus muntjak* (muntjac). Regarding angiosperm plants: *Actinidia deliciosa* (kiwifruit), *Arabidopsis thaliana* (thale cress) and *Viscum album* (common mistletoe). We have excluded the point of the plant *Voanioala gerardii*, with $L_g = 298$, $L_c \approx 98.6$, for the sake of clarity.

Group	N	ρ	p
fungi	54	-0.556	< 0.001
angiosperm plants	3408	-0.454	< 0.001
gymnosperm plants	171	0.124	0.107
insects	210	-0.442	< 0.001
reptiles	54	-0.331	< 0.001
birds	99	-0.558	< 0.001
mammals	373	-0.782	< 0.001
cartilaginous fishes	52	-0.463	0.001
jawless fishes	13	-0.871	< 0.001
ray-finned fishes	621	0.075	0.063
amphibians	316	0.270	< 0.001
ray-finned fishes from the genus *Oncorhynchus*	11	-0.641	0.034
ray-finned fishes from the genus *Corydoras*	18	-0.611	0.007
gymnosperm plants excluding the genus *Pinus*	106	-0.238	0.014

Table 1

Correlations between genome size and chromosome length. We define N, ρ and p, respectively, as the number of different organisms, the value of Spearman's rank correlation statistic for L_g versus L_c and p as the p-value of ρ within a group of organisms. Lobe-finned fishes were excluded from our study because there were not enough species ($N = 5$).